Linked Data in Production
Moving Beyond Ontologies

David Newbury
Assistant Director, Software and User Experience, Getty

CNI Spring Project Briefings, March 26, 2024
Thank you for inviting me to present today.

I am a technologist working in the cultural heritage sector.

I lead the public digital team, developing applications to support Getty’s mission.
I work for Getty in Los Angeles, which is a library/archive/museum/research center.

One of our major areas of digital leadership is in the use of Linked Data for cultural heritage.
Linked Open Data is a set of technologies that attempt to translate some of the best practices of the Web for use with structured data:

- The use of URLs as identifiers
- Networks of information, not tables
- Formal, shared standards for description

Introduction: **What is Linked Open Data?**

[Diagram showing a network of interconnected circles labeled with various terms related to Linked Open Data and related projects.]
Introduction: **Linked Data is (mostly) Dead.**

As it turns out, Linked Data is not wildly successful.

But we all still talk about it a lot.
Linked Data’s appeal in cultural heritage is a technological solution to a social problem:

Cheap storage, ubiquitous connectivity, and search algorithms recontextualize the labor behind cultural heritage data work.
Mass digitization, computational metadata generation, and decades of cataloguing mean that our institutions have more data to provide than we have the ability to provide context for.

Data overload and limited user attention are the collections access problems of the next decade.
Our always-connected culture means that our collections are increasingly seen as part of a single digital ecosystem—

And the questions that are being asked require information that extends beyond the boundaries of any one institution.
And commercial tools have given users the expectation that information is available for the asking—

Eliding the labor and capital needed to create, curate, and maintain that information.
Linked Data has been seen as a solution:

It provides structures that manage the scale of data we create,

identifiers that maintain authority in a globally distributed environment,

and ontologies that enable complex data retrieval across datasets.
What came before:

Laying the Groundwork
In 2014, the Getty Vocabularies were launched as Linked Data. This, alongside the work at Yale Center for British Art, Rijksmuseum, and the British Museum, demonstrated the feasibility of LOD within the museum community.
The archives of the Carnegie Museum of Art’s Film Department launched in 2014.

The animating question was:

What would happen if you treated the **relationships** between events, archival material, people, and artwork as the essential element, **not the objects**?
In 2017, the American Art Collaborative launched.

It used these same principles to highlight connections across 14 institutions and 152,000 items—using Getty’s Vocabularies as a bridging structure between institutions.
One of the most lasting outcomes of the American Art Collaborative was **Linked.Art**, the shared data model that connected institutions.

---

Laying the Groundwork: **Linked Art**

Digital Objects

All Digital Objects share some basic characteristics, regardless of their particular nature. The basic patterns of **Name**, **Identifier**, **Classification**, and **Statement** all apply in the regular way. Beyond the baseline, digital objects can have the following descriptive features:

- **Access Point** - The URLs where the object is available. These may also treated as Digital Objects in their own right, but typically only the URL is given.
- **Format** - The format of a digital object is its media type, often called a MIME type, given as a string.
- **Standard** - Many digital objects further conform to standard specifications, which can be referenced using the **conforms_to** property. This differs from **format**, as there may not be a media type for the specification, and from **classified_as**, which is a broader classification (image, rather than conforming to the standard for JPEG 2000).
- **Dimensions** - Digital dimensions follow the same pattern as **physical_dimensions**, but might use different types (file size) or the same (height, width for images) and different units (bytes, pixels).
- **Creation** - Digital Objects are created by **Creation events** rather than **Production events**, but otherwise have the same activity model.

**Example:**

A web page was created by a Museum Education department using the HTML format (and standard), is 100k in size, and is available on the museum website.

```
{
  "content": "https://linked-art.org/","id": "https://linked-art/example/digital/1","type": "DigitalObject","_label": "Digital Object"},
"classified_as": [ 
  { 
    "id": "http://vocab.getty.edu/aat/302064708",
    "type": "Type",
    "_label": "Web Page"
  }
],
"identified_by": [ 
  { 
    "type": "Name",
    "content": "Informative Web Page"
  }
],
"dimension": [ 
  { 
    "type": "Dimension",
    "_label": "1MB kb",
    "classified_as": [ 
      { 
        "id": "http://vocab.getty.edu/aat/302065003",
        "type": "Type",
        "_label": "file Size"
      }
    ]
  }
],
"value": 100,
"unit": [ 
  { 
    "id": "http://vocab.getty.edu/aat/302065070",
    "type": "MeasurementUnit",
    "_label": "100kb"}
]"}
What we’ve done:

Getty’s Digital Ecosystem
Getty has been doing Linked Data since 2014, starting with the Getty Vocabularies.

It’s a thesaurus of concepts, people, and places used for cataloging across many institutions.
Since then, we’ve moved most of our major systems to use Linked Data—including our archives...
Since then, we’ve moved most of our major systems to use Linked Data—including our archives...

... and our museum collection.
It’s used for onsite visitor experiences via our audio guide...
It’s used for onsite visitor experiences via our audio guide...

...and to provide novel interfaces for exploration of our materials.
It’s also used by third parties: both large, like Google Arts & Culture...
It’s also used by third parties: both large, like Google Arts & Culture...

...and small, like this project by the Cultural Office of the Embassy of Spain.
We’ve also built a complex, powerful digital infrastructure to support this work—millions of records in a single shared data model, pulling from a wide collection of systems of record.
Under the hood:
Linked Data & the Everything API
Data Flow: How we wish it was

Staff Interface

Public Website
These systems support people.

Digital infrastructure is designed to use computers to empower people to be more effective at meeting the mission of the organization.
Catalogers need systems that match their workflows—and different disciplines have different needs.

Our infrastructure needed to not be tied to any particular backend system.
Engineers need to get data in and out of systems, using patterns and practices that they already know how to use.
System admins just don’t want you to break their stuff.

Pulling data out of systems on demand usually breaks stuff.
Most end users are looking for content—they want to learn what we know on a given topic.

This may be professional scholarship or it might be looking for pictures—both are examples of information-seeking behaviours.
And some researchers want to find questions that haven’t been asked before—to find new connections or patterns in the data that others have overlooked.
Meeting the needs of catalogers is mostly not my problem.

There are high-quality, professional tools that work within the disciplinary training of the field.
Providing access to that data, though, often requires recontextualization:

Changing the conceptual lens from one focused on staff efficiencies to one focused on user’s needs.
Doing so requires combining data from multiple systems and multiple workflows into a new record.

This combining—or linking—of data has tradeoffs.
Imagine a record for the painting *Irises*.
And a second record, this one for Van Gogh.
These could be seen as two separate documents:

```json
"@context": "https://linked.art/ns/v1/linked-art.json",
"id": "person/1",
"type": "Person",
"identified_by": {
  "id": "person/1/name",
  "type": "Name",
  "content": "Vincent Van Gogh"
}
```

```json
"@context": "https://linked.art/ns/v1/linked-art.json",
"id": "object/1",
"type": "HumanMadeObject",
"identified_by": {
  "id": "object/1/name",
  "type": "Name",
  "content": "Irises"
},
"produced_by": {
  "id": "object/1/production",
  "carried_out_by": {"id": "person/1"}
}
```
Or as a single graph.

Getty's Linked Data: The LOD Gateway
From the point of view of the data, these are equivalent—they contain the same facts.

But from a usability perspective, they make different things easy or hard.
Documents: For Access and Discovery

Documents are optimized for **Access:**

They provide a specific set of data bundled together by the data creator that provide all the facts you need...given a **specific context.**
Graphs: *For Queries*

Graphs, alternately, are optimized for **querying:**

Allowing a user to define a specific context based on novel criteria and returning that subset of facts.
Imagine two Questions:

“What objects does Getty have that have images larger than 1200px on the longest side that have been exhibited in both New York and Paris and were created by artists who lived before 1850? and

What’s the label info for *Irises*?
Imagine two Questions:

At the Getty, we have never asked:

“What objects does Getty have that have images larger than 1200px on the longest side that have been exhibited in both New York and Paris and were created by artists who lived before 1850?

but we ask

What’s the label info for *Irises*?

Several thousand times a day.
Having an interface for documents lets us provide a simple, easily understandable record that maps well to known contexts. This is important, because people usually expect these contexts. It makes answering common questions simple.
Documents are also the way the internet works: REST APIs, cache control, JSON, webpages.

Using these well-known systems helps developers make systems that are fast and easy to build.
Research is different—each scholar brings their own question and their own context.

Meeting their need means empowering them to draw their own boundaries within the data.
Doing so is complex—it moves the burden of defining the relevant context to the user of the data, not the creator of the data.

But it makes asking new questions possible, even if it might be inefficient or complicated.
Meeting Both Needs

We’ve built our infrastructure to allow for both use cases:

A developer can create, update, and delete documents, and behind the scenes it will keep a graph in sync with those changes.
Linked Data Infrastructure: **Tracking Changes**

It also allows for synchronization across systems:

A editor changes a record, which means the API needs updated, which means the website needs updated, and the search interfaces, and third-party systems...
The infrastructure uses the **W3C ActivityStream** standard and are implemented using the patterns from the IIIF Change Discovery API.

Using standards makes it easy to build integrations against changing data—both within our organization and for external aggregators.
For some kinds of data, it's also valuable to also be able to see **what has changed over time** for a given record.

To do so, our APIs also supports **Memento**, the standard underneath the Internet Archive.
This lets you automatically open older versions of the record—providing an audit log and the ability for scholars to understand how knowledge changes over time.
Cool Tech, Bro:

Why Does this Matter?
A Hard-won lesson:

No application that we’ve built **required** Linked Data.
A Hard-won lesson:

No application that we’ve built **required** Linked Data.

Which, if you think about it, makes sense. Each application has a **specific, known context** with clear record boundaries.
A Hard-won lesson:

Different users have different contexts and need different affordances.

A shared, graph-based data model allows us to re-present the data in a way that matches user’s varying models of the world via multiple interfaces.
As Simple as possible:

A shared data model also makes our developers more effective—eventually.

Building on top of web technology lets the engineering learning curve be gradual.
As Simple as possible:

Standards are valuable for interoperability—but also because you don’t have to write all the documentation.

Nobody wants to write it, but you can’t work across institutions without it.
As Simple as possible:

Minimize complexity in the data model.

Data is for computers—text is for humans. Resist the urge to show off.

You can always add complexity—you can never take it away.
Disciplinary Misdeeds:

The hardest part of this will be change management.

Recontextualizing information across boundaries hides disciplinary labor—and digital innovations can conflict with pre-digital best practices.
Evangelize and collaborate.

What makes cultural data interesting is not contained within any one institution.

It’s shared across our entire, world-wide community. We should work together.

Shared models and shared code make that easier.
Over the next several years, we’ll be expanding our usage of this system:

This fall, we’ll launch a new version of the Getty Provenance Index, adding in 22M records of transactions between art dealers.

This research-focused dataset will allow new insights into collections around the world—and into the art market as a whole.
We’re beginning to plan the next iteration of the Getty Vocabularies infrastructure:

Working to understand how the multiple contexts of our audiences can be supported—and how new ways of working impact the platform.
And we’re using the **platforms** and **standards** we’ve put in place to enable collaboration across the field:

Working with the Smithsonian to provide **joint access and discovery** for millions of images from the photo morgue of magazines such as Ebony and Jet.
Why do we do Linked Data?

The value is not in the technologies or the ontologies we use.
Why do we do Linked Data?

The value is in the **ecosystem**—information in varied context for different applications.

The value is in the **audience**—supporting user needs and conceptual models.

And it’s in the **community**—allowing data and code to be used beyond the Getty.
Why do we do Linked Data?

We do it for humans.
Thank You.

Contact me at dnewbury@getty.edu

David Newbury
Assistant Director, Software and User Experience, Getty

CNI Spring Project Briefings, March 26, 2024